Biometrics 74, 1055–1064
September 2018

DOI: 10.1111/biom.12864

Bayesian Enhancement Two-Stage Design for Single-Arm Phase II Clinical Trials with Binary and Time-to-Event Endpoints

Haolun Shi* **and Guosheng Yin****

Department of Statistics and Actuarial Science, The University of Hong Kong, Hong Kong
**email:* shl2003@connect.hku.hk
***email:* gyin@hku.hk

Summary. Simon's two-stage design is one of the most commonly used methods in phase II clinical trials with binary endpoints. The design tests the null hypothesis that the response rate is less than an uninteresting level, versus the alternative hypothesis that the response rate is greater than a desirable target level. From a Bayesian perspective, we compute the posterior probabilities of the null and alternative hypotheses given that a promising result is declared in Simon's design. Our study reveals that because the frequentist hypothesis testing framework places its focus on the null hypothesis, a potentially efficacious treatment identified by rejecting the null under Simon's design could have only less than 10% posterior probability of attaining the desirable target level. Due to the indifference region between the null and alternative, rejecting the null does not necessarily mean that the drug achieves the desirable response level. To clarify such ambiguity, we propose a Bayesian enhancement two-stage (BET) design, which guarantees a high posterior probability of the response rate reaching the target level, while allowing for early termination and sample size saving in case that the drug's response rate is smaller than the clinically uninteresting level. Moreover, the BET design can be naturally adapted to accommodate survival endpoints. We conduct extensive simulation studies to examine the empirical performance of our design and present two trial examples as applications.

Key words: Bayesian adaptive design; Highest posterior density interval; Phase II clinical trial; Posterior probability; Simon's design; Survival endpoint.

1. Introduction

Recently, a multicenter, multinational, double-blind, randomized phase III clinical trial comparing the efficacy of erlotinib plus the standard drug (sorafenib) versus the standard drug alone in patients with advanced hepatocellular carcinoma (Zhu et al., 2015) concluded with a failure. A total of 720 patients were randomized and the primary endpoint was the overall survival time. At the end of the study, the overall survival, the progression-free survival, and the overall response rate all failed to achieve statistically significant treatment differences. Such a large-scale phase III trial was preceded by a single-arm phase II study on erlotinib (Philip et al., 2005), where the drug was declared as promising for further studies. The primary endpoint of this phase II trial was binary, which equaled 1 if the patient was progression-free at 24 weeks and 0 otherwise. The phase II study adopted Simon's two-stage design, with a specified uninteresting null response rate of 5% and a target response rate of 20%.

The failures of large-scale phase III studies with promising drugs identified in phase II trials are not uncommon. In fact, studies revealed that approximately 45% of all drugs that entered the phase III programs ended with statistically insignificant results, and in oncology, this percentage is as high as 59% (Kola and Landis, 2004). A more recent report showed that 62% of the 235 phase III randomized cancer trials published in 10 journals between 2005 and 2009 failed to achieve significant results (Gan et al., 2012). Considering the substantial amount of time and resources consumed when conducting a phase III study, such a failure percentage is considered as unacceptably high. In reality, a large proportion of phase II "promising" drugs unfortunately fail in phase III studies, which casts doubt on the reliability of existing phase II methods. There are studies that reflected on the underlying reasons for such a high failure rate and called for an improvement in the efficiency of phase II trial designs (Retzios, 2009). It is critical to obtain more adequate information and make a more sound go/no-go decision at the end of phase II trials, which would improve the success rate in phase III trials and achieve substantial cost saving (Yin, 2012).

Single-arm trials are often preferred and utilized in phase II studies due to the low sample size requirement and the adaptive feature of futility stopping. Simon (1989) proposed a hypothesis testing framework that fulfills the requirements on the type I and type II error rates. Green and Dahlberg (1992) developed designs when the attained sample size varies from the originally planned one. Ensign et al. (1994) extended Simon's two-stage design to the three-stage setting. Shuster (2002) proposed a two-stage design that allows the efficacy stopping rule and aimed at minimizing the maximum expected sample size over a group of response rates. Similarly, Chen and Shan (2008) developed optimal and minimax three-stage designs with efficacy stopping rules. Koyama and Chen (2008) studied how to draw proper inference under Simon's two-stage design. Mander and Thompson (2010) considered optimizing the expected sample size in Simon's design based on the alternative hypothesis when the efficacy stopping rule is allowed. Liu et al. (2010) extended Simon's design based on beta–binomial distributions and derived its asymptotic properties. Baey and Le Deley (2011) studied the effects of the misspecification of the prespecified response rates in Simon's design. Shan et al. (2016) proposed a flexible modification of Simon's design by allowing the second-stage sample size to depend on the first-stage response rate, which leads to a smaller expected sample size compared with Simon's optimal design. Along the line, phase II trial designs have also been developed extensively under the Bayesian paradigm. Thall and Simon (1994) proposed a Bayesian single-arm design for phase II trials that continuously monitors the binary outcomes and makes adaptive decisions based on posterior probabilities, and Lee and Liu (2008) extended the design using predictive probabilities of future possible outcomes at the end of the trial. Tan and Machin (2002) proposed Bayesian single and dual threshold designs, and Sambucini (2010) developed a two-stage Bayesian predictive method which adjusts the sample size based on the observed data in the first stage.

For single-arm phase II trials, the two-stage design developed by Simon (1989) is the most commonly used approach, which has been adopted by over 20% of all phase II studies (Lee and Feng, 2005) and cited over 2000 times. From a Bayesian perspective, we study the posterior probabilities of the null and alternative hypotheses given that a promising result is declared in Simon's two-stage design. Our study reveals that in some cases, a potentially efficacious treatment identified by Simon's design could have only less than 10% posterior probability of actually reaching the desirable target level. We consider such a low probability of achieving the minimum efficacy level as a potential culprit for the high failure rate among those seemingly promising drugs that are carried forward from single-arm phase II trials into phase III studies. To address this issue, we propose the Bayesian enhancement two-stage (BET) design, where both the first and the second stages base decisions on the posterior probabilities: the first stage compares the drug's response rate with a clinically uninteresting rate for futility stopping, and the second stage compares it with the desirable target level for declaring efficacy. Compared with Simon's design, the BET design guarantees a high posterior probability of the response rate reaching the target level, while allowing for trial early termination and sample size saving in case the drug's response rate is lower than the clinically uninteresting level.

The article is organized as follows. In Section 2, we present an analysis based on the posterior probabilities under Simon's two-stage design, and propose the BET designs for binary and survival endpoints. Section 3 describes the simulation studies of the BET designs and Section 4 presents two trial examples. Finally, Section 5 concludes the article with some remarks.

2. BET Design

2.1. *Simon's Two-Stage Design*

Simon (1989) proposed a single-arm two-stage design based on the hypotheses $H_0: p \leq p_0$ versus $H_1: p \geq p_1$, where p_0 denotes a clinically uninteresting response rate, p_1 represents the desirable target response rate, and the gap between p_0 and p_1 is called the indifference region. Simon's design is characterized by four parameters (n_1, n, r_1, r). Let y_1 and y_2 denote the number of responses observed in the first and second stage, respectively. The first stage sample size is n_1, and if $y_1 \geq r_1$, the trial would proceed into the second stage; otherwise, the trial is terminated early for futility. In the second stage, an additional sample size of $n_2 = n - n_1$ is enrolled, and if the total number of responses $y_1 + y_2$ reaches r, the null hypothesis H_0 is rejected and the drug is declared as promising; otherwise, the drug is considered as nonpromising. It is worth emphasizing that the notations r_1 and r here differ from those in the original article by Simon (1989), that is, in our article, the drug is declared as futile if $y_1 < r_1$, whereas using the same notation in Simon (1989), the trial stops for futility if $y_1 \leq r_1$. Subtracting one from the threshold values in our design would result in the same interpretation as in Simon's original design. The design parameters are calibrated under the constraints on type I and type II error rates, denoted as α and β, respectively. Among all the admissible design parameters that satisfy such constraints, we may choose the one with the smallest expected sample size when $p = p_0$ (the "optimal" design), or the one with the smallest total sample size (the "minimax" design).

It should be cautioned that in Simon's two-stage design rejecting the null hypothesis H_0 does not imply accepting the alternative H_1 due to the indifference region between p_0 and p_1. In fact, Simon's design only warrants good confidence in concluding that the drug's response rate is larger than p_0, the clinically uninteresting rate. Often, the data corresponding to the design parameters in Simon's two-stage design often fall in the indifference region (close to the middle of p_0 and p_1), that is, the data would support rejection of H_0 as well as rejection of H_1. The frequentist hypothesis testing framework only examines whether the data would be rare under the null, but does not consider whether the data would also be rare under the alternative.

A more straightforward way to understand such an inherent problem is to look at the posterior probabilities of the null and alternative hypotheses when the response number equals r_1 and r at the end of the first stage and the final stage. Table 1 presents Simon's optimal designs under various common specifications of (p_0, p_1) and constraints of (α, β). Assuming a uniform prior on p, that is, $p \sim \text{Beta}(1, 1)$, the posterior distribution of p also follows a beta distribution due to the conjugacy property between a beta prior and a binomial likelihood. As a result, it is easy to compute the posterior probabilities of H_0 and H_1. For example, $\Pr(H_0|r_1, n_1)$ denotes the posterior probability of $p \leq p_0$ when the response number among n_1 patients reaches r_1, the minimum required level for continuation at the end of the first stage, and $\Pr(H_1|r, n)$ denotes that of $p \geq p_1$ when the total response number among n patients is equal to r, the minimum required level for declaring the drug promising at the end of the whole trial. Because Simon's optimal two-stage design is more focused on p_0, under the scenarios where the effectiveness of the drug is declared, the posterior probability of H_0 is close to or smaller than α. However, when the drug is declared as promising at the end of the second stage, or when the continuation criterion is met at the end of the first stage, the posterior

Table 1

Simon's optimal two-stage designs under various specifications of (p_0, p_1) and (α, β), and the posterior probabilities of H_0 and H_1 when the response numbers reach the minimum required levels at the end of the first and second stages, respectively

p_0	p_1	α	β	n_1	n	r_1	r	$\Pr(H_0 \mid r_1, n_1)$	$\Pr(H_1 \mid r_1, n_1)$	$\Pr(H_0 \mid r, n)$	$\Pr(H_1 \mid r, n)$
0.05	0.25	0.1	0.1	9	24	1	3	0.0861	0.2440	0.0341	0.0962
		0.1	0.2	6	23	1	3	0.0444	0.4449	0.0298	0.1150
		0.2	0.2	6	16	1	2	0.0444	0.4449	0.0503	0.1637
0.1	0.3	0.1	0.1	12	35	2	6	0.1339	0.2025	0.0628	0.0536
		0.1	0.2	7	18	1	4	0.1869	0.2553	0.0352	0.2822
		0.2	0.2	6	15	1	3	0.1497	0.3294	0.0684	0.2459
0.2	0.4	0.1	0.1	17	37	4	11	0.2836	0.0942	0.0623	0.1089
		0.1	0.2	12	25	3	8	0.2527	0.1686	0.0592	0.2255
		0.2	0.2	11	16	3	5	0.2054	0.2253	0.1057	0.2639
0.3	0.5	0.1	0.1	22	46	8	18	0.2291	0.1050	0.0831	0.0719
		0.1	0.2	15	32	6	13	0.1753	0.2272	0.0884	0.1481
		0.2	0.2	6	20	2	8	0.3529	0.2266	0.1477	0.1917

probability of H_1 being true is also very small. It is evident that the column showing $\Pr(H_1|r, n)$ indicates that the posterior probability of the response probability reaching the target desired level is in fact very low, ranging from 5% to 28%. As a conclusion, rejection of H_0 under Simon's two-stage design only indicates that the drug cannot be claimed to be futile but still it cannot be claimed to be promising or clinically meaningful, because the observed data often fall inside the indifference region that shows little support to either H_0 or H_1.

Alternatively, we can gain more insight by examining the density curve of the posterior distribution. The upper panel of Figure 1 displays the posterior distributions when the response number reaches r at the final stage, under a range of values of p_0 and p_1 (the distance between p_0 and p_1 becomes smaller). It is clear that the modes of the posterior distributions always fall halfway between p_0 and p_1 regardless of how close p_0 and p_1 are. In other words, when Simon's design claims a drug to promising, the observed response rate in fact falls in the middle of the indifference region, so that the collected information does not support H_0, while it does not support H_1 either.

2.2. *BET Design for Binary Endpoint*

To impose a more stringent efficacy evaluation in single-arm phase II trials, we propose a Bayesian enhancement two-stage (BET) design based on posterior probabilities of H_0 and H_1. In the first stage, n_1 subjects are enrolled, and suppose that we observe y_1 responses, then $y_1|p \sim \text{Bin}(n_1, p)$, where $\text{Bin}(n, p)$ denotes the binomial distribution with a success probability p. The response rate of the experimental drug is assumed to follow a beta prior distribution, $p \sim \text{Beta}(a, b)$. If there is little information on the efficacy of the experimental drug, we may take the uniform prior distribution for p with $a = b = 1$. By the conjugacy property of the beta distribution, the posterior distribution of p at the end of the first stage is $\text{Beta}(a + y_1, b + n_1 - y_1)$. Based on the posterior distribution, we compute the posterior probability of the experimental response rate being greater than p_0 as

$$\text{PoP}_1 \equiv \Pr(p > p_0|y_1, n_1) = \int_{p_0}^{1} \frac{p^{a+y_1-1}(1-p)^{b+n_1-y_1-1}}{B(a+y_1, b+n_1-y_1)} dp,$$

where $B(a, b) = \Gamma(a)\Gamma(b)/\Gamma(a + b)$ denotes the beta function with parameters a and b.

At the end of the first stage, if $\text{PoP}_1 > \pi_1$, where π_1 is a prespecified cutoff for the posterior probability, the trial proceeds into the second stage; otherwise, the trial is terminated for futility. In the second stage, additional $n - n_1$ subjects are enrolled and let y_2 be the number of responses among these subjects. The posterior distribution of p at the end of the second stage is $\text{Beta}(a + y_1 + y_2, b + n - y_1 - y_2)$, and the posterior probability of the experimental response rate being greater than p_1 is

$$\text{PoP}_2 \equiv \Pr(p > p_1|y_1 + y_2, n) = \int_{p_1}^{1} \frac{p^{a+y_1+y_2-1}(1-p)^{b+n-y_1-y_2-1}}{B(a+y_1+y_2, b+n-y_1-y_2)} dp.$$

We claim the experimental drug promising if $\text{PoP}_2 > \pi_2$; otherwise, we declare the drug nonpromising. Although the sampling distribution of $y = y_1 + y_2$ is conditional on the event $y_1 \geq r_1$, such a condition regarding the response number and the stopping rule at the end of the first stage does not affect the beta posterior distribution of p in the second stage (Sambucini, 2008).

To determine the sample sizes and the minimum required numbers of responses at the end of the first and second stages, respectively, we define the following criteria based on the lengths of the highest posterior density (HPD) intervals. Given a response number y and the sample size n, an HPD interval on the posterior distribution $\text{Beta}(a + y, b + n - y)$ can be obtained for p with a coverage probability of π, and let $l_p(\pi|y, n)$ denote its length. The design parameters

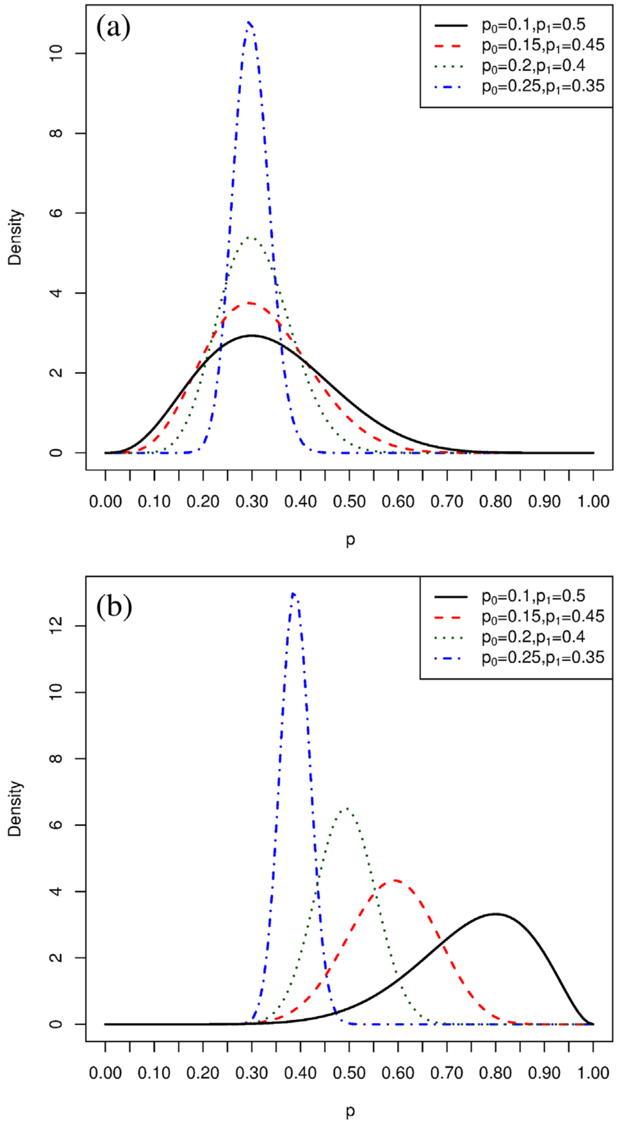

Figure 1. Comparison of posterior distributions when the total response number reaches the minimum required level over a range of specified (p_0, p_1) under (a) Simon's optimal two-stage designs and (b) Bayesian enhancement two-stage designs.

(n_1, n, r_1, r) should satisfy

$$l_p(\pi_1|r_1, n_1) < \ell_1, \quad \Pr(p > p_0|r_1, n_1) > \pi_1,$$
$$l_p(\pi_2|r, n) < \ell_2, \quad \Pr(p > p_1|r, n) > \pi_2,$$

where ℓ_1 and ℓ_2 are the desirable lengths of the HPD intervals at stage 1 and stage 2, respectively. The values of ℓ_1 and ℓ_2 can be determined with consideration of the sample size constraint. As the cost of falsely stopping for futility is smaller than that of falsely continuing the trial with an inefficacious drug into further phase III studies, we typically set $\pi_2 \geq \pi_1$. The first two inequalities are used for solving (n_1, r_1) and the remaining two for (n, r). Under a given sample size n, the minimum required number of responses is pinpointed first, as the posterior probability is a monotonic function of r. The length of the HPD interval is then computed based on n and r. We enumerate the sample size (n_1 or n) until the corresponding minimum required number of responses (r_1 or r) is associated with an adequately narrow HPD interval.

The searching algorithm for the optimal parameters (n_1, r_1) is described as follows.

(1) We start with $n_1 = n_{\min}$, the minimum sample size in the first stage, and typically $n_{\min} = 1$.
(2) Given $n_1 = i$, find the minimum number of responses y_1 such that $\Pr(p > p_0|y_1, n_1) > \pi_1$, and calculate the corresponding HPD interval length $l_p(\pi_1|y_1, n_1)$.
(3) If $l_p(\pi_1|y_1, n_1) < \ell_1$, we stop the algorithm and set $n_1 = i$ and $r_1 = y_1$; otherwise, we set $n_1 = i + 1$ and repeat step (2).

The algorithm for (n, r) can be developed along similar lines by substituting ℓ_1 for ℓ_2, π_1 for π_2, and p_0 for p_1.

2.3. *BET Design for Survival Endpoint*

While it is difficult to develop the counterpart of Simon's two-stage design for survival endpoint, our BET design can be naturally adapted to accommodate time-to-event data. We denote T as the time of failure, and C as the time of censoring. The observed event time is $Y = \min(T, C)$, with a censoring indicator $\Delta = I(T \leq C)$, where $I(\cdot)$ is the indicator function.

The survival time is assumed to follow a Weibull distribution, with a density function

$$f(t|k, \lambda) = \frac{k}{\lambda} t^{k-1} \exp\left(-\frac{t^k}{\lambda}\right), \quad k > 0, \lambda > 0.$$

Let $\theta = (\lambda \ln 2)^{1/k}$ denote the median of the Weibull distribution. We are interested in testing hypotheses on the median survival time θ, $H_0 : \theta \leq \theta_0$ versus $H_1 : \theta \geq \theta_1$, where θ_0 is the maximum median survival time for a futile drug, and $\theta_1 > \theta_0$ is the minimum median survival time for an efficacious drug.

We specify a diffuse exponential prior distribution for k, that is, $k \sim \text{Exp}(b_k)$ with $b_k = 0.01$, and a diffuse inverse gamma prior distribution for λ, that is, $\lambda \sim \text{IGamma}(a_\lambda, b_\lambda)$ with $a_\lambda = b_\lambda = 0.01$, whose density functions are denoted by $p(k)$ and $p(\lambda)$, respectively. Let $D = \{(y_i, \Delta_i), i = 1, \ldots, n\}$ denote the observed data, where n is the cumulative sample size. The joint posterior distribution of k and λ is given by

$$p(k, \lambda|D) = \left(\frac{k}{\lambda}\right)^{\sum_{i=1}^n \Delta_i} \left(\prod_{i=1}^n y_i^{\Delta_i(k-1)}\right) \exp\left(-\frac{\sum_{i=1}^n y_i^k}{\lambda}\right) \times p(k) p(\lambda).$$

Due to the conjugacy relationship between the inverse gamma distribution and the Weibull likelihood function, the full conditional distribution of λ also follows an inverse gamma distribution,

$$\lambda|k, D \sim \text{IGamma}\left(\sum_{i=1}^n \Delta_i + a_\lambda, \sum_{i=1}^n y_i^k + b_\lambda\right).$$

We denote $d = \sum_{i=1}^{n} \Delta_i$ and $\bar{y} = \sum_{i=1}^{n} y_i^k / n$, and denote the full conditional distribution of λ as $p(\lambda | k, d, \bar{y})$. The full conditional distributions of k is given by

$$p(k|\lambda, D) \propto k^{\sum_{i=1}^{n} \Delta_i} \left(\prod_{i=1}^{n} {y_i}^{\Delta_i (k-1)} \right) \exp \left(- \frac{\sum_{i=1}^{n} {y_i}^k}{\lambda} - kb_k \right).$$

To determine the stopping boundaries for the first and second stages, we define the following criteria based on the lengths of the HPD intervals. Let k_{HPD} denote a prespecified value of k for calculating the HPD interval. The length of the HPD interval of θ given $k = k_{\text{HPD}}$ with a coverage probability of π_1 is $l_\theta(\pi_1 | k_{\text{HPD}}, d, \bar{y}) = (\lambda_U \ln 2)^{1/k_{\text{HPD}}} - (\lambda_L \ln 2)^{1/k_{\text{HPD}}}$, where λ_L and λ_U denote the lower and upper boundaries of the HPD interval of λ based on $p(\lambda | k_{\text{HPD}}, d, \bar{y})$, respectively. It is worth emphasizing that k_{HPD} is only used for the calculation of the HPD interval and thus the sample size, whereas k is unknown and needs to be estimated.

During the first stage, upon the observation of a failure event, we check whether the HPD interval length satisfies

$$l_\theta(\pi_1 | k_{\text{HPD}}, d_1, m_1) < \ell_1, \tag{2.1}$$

where ℓ_1 is the prespecified desirable length of the HPD interval for the first stage, and $d_1 = \sum_{i=1}^{n_1} \Delta_i$ is the cumulative number of events. Moreover, m_1 can be understood as the minimum value of $\bar{y}_1 = \sum_{i=1}^{n_1} y_i^k / n_1$ that corresponds to the posterior probability of $\theta > \theta_0$ being no less than π_1; that is,

$$\Pr(\theta > \theta_0 | k_{\text{HPD}}, d_1, m_1) = \pi_1,$$

based on $\theta = (\lambda \ln 2)^{1/k_{\text{HPD}}}$ and the conditional distribution $p(\lambda | k_{\text{HPD}}, d_1, \bar{y}_1)$. If (2.1) is not satisfied, we extend the follow-up time and continue to recruit more subjects. If (2.1) is satisfied, the trial has acquired sufficient amount of information for decision making: if the posterior probability of $\theta > \theta_0$ exceeds π_1, that is, $\text{PoP}_1 \equiv \Pr(\theta > \theta_0 | D_1) > \pi_1$, where D_1 denotes the observed data in the first stage, the trial would proceed into the second stage; otherwise, the trial would be stopped early for futility.

In the second stage, we continue to recruit more subjects, and upon the observation of a failure event, we check whether the data are adequate for decision making based on the HPD interval length,

$$l_\theta(\pi_2 | k_{\text{HPD}}, d, m) < \ell_2, \tag{2.2}$$

where ℓ_2 is the prespecified desirable length of the HPD interval for the second stage. Moreover, m can be understood as the minimum value of $\bar{y}$ that corresponds to the posterior probability of $\theta > \theta_1$ being no less than π_2; that is,

$$\Pr(\theta > \theta_1 | k_{\text{HPD}}, d, m) = \pi_2,$$

based on $\theta = (\lambda \ln 2)^{1/k_{\text{HPD}}}$ and the conditional distribution $p(\lambda | k_{\text{HPD}}, d, \bar{y})$. If (2.2) is not satisfied, we extend the follow-up time and recruit more subjects; otherwise, we terminate the trial for a conclusion. As a final decision, we declare the drug as promising if the posterior probability of $\theta > \theta_1$ exceeds π_2, that is, $\text{PoP}_2 \equiv \Pr(\theta > \theta_1 | D) > \pi_2$, and unpromising otherwise.

Due to the complexity caused by censoring, instead of prespecifying fixed sample sizes, we let the data determine whether the trial should be stopped for decision making. The HPD interval length serves as an indicator of the amount of information that the trial has accumulated. Unlike the binary data case, the BET design with survival endpoint does not search for the minimum sample size that satisfies certain constraint on the posterior distribution, due to the uncertainty caused by censoring. The BET design with survival endpoint also consists of an interim and a final analysis. However, the times when such analyses are performed cannot be prefixed but determined by the accumulated data in the trial. This allows the trial conduct to adapt flexibly in accord with the censoring percentage: the higher the censoring percentage, the more samples required for delivering a conclusive decision.

The motivation behind computing the HPD interval under a prefixed value of $k = k_{\text{HPD}}$ instead of the actual HPD interval from the data is to improve the stability and manageability of the design characteristics. With a fixed $k = k_{\text{HPD}}$, the HPD interval length is less variable and has a decreasing relationship with the accumulated sample size. Moreover, such a method reduces the computational burden as there is no need to conduct Markov Chain Monte Carlo every time a failure event is observed, but only after the HPD interval length is short enough. As the design parameters (ℓ_1, ℓ_2) need to be calibrated based on the sample size constraints, the actual value of k_{HPD} is of minor importance; for example, we may set $k_{\text{HPD}} = 1$ as default.

3. Simulation Studies

We conduct extensive simulation studies to examine the performance of the proposed BET design. We assess paired values of (p_0, p_1) where p_0 ranges from 0.05 to 0.3, and the effect size (i.e., the difference between p_1 and p_0) takes a value of 0.2. For each pair of (p_0, p_1), we calibrate the design parameters (n_1, n, r_1, r) based on three sets of the desired HPD interval length, $(\ell_1, \ell_2) = (0.25, 0.2)$, $(0.3, 0.22)$, and $(0.35, 0.25)$. We take a noninformative prior distribution $p \sim \text{Beta}(1, 1)$, and set $(\pi_1, \pi_2) = (0.8, 0.9)$. Table 2 shows the solutions of (n_1, n, r_1, r) under different trial specifications. In addition, we compute the posterior probability of H_0 and H_1 when r_1 responses are observed in stage 1, and when a total of r responses are observed at the end of stage 2, respectively. Clearly, at the end of stage 1, if the response number meets the continuation criterion, the BET design ensures that the posterior probability of H_0 being true is smaller than $1 - \pi_1 = 0.2$. If the drug is declared as promising in stage 2, the posterior probability of H_1 being true is greater than $\pi_2 = 0.9$, and the posterior probability of H_0 being true becomes very close to zero. As expected, the sample size decreases when the length of the HPD interval increases. Comparing the posterior probabilities under Simon's optimal two-stage design in Table 1 with those under the proposed design, it is evident that the BET design is more directed toward the effectiveness of the drug by demonstrating the response rate to be greater than

Table 2

Bayesian enhancement two-stage designs for binary endpoint with $(\pi_1, \pi_2) = (0.8, 0.9)$ under Beta(1,1) and Beta(8,12) prior distributions and various specifications of (p_0, p_1) and (ℓ_1, ℓ_2), and the posterior probabilities of H_0 and H_1 when the response numbers reach the minimum required levels at the end of the first and second stages, respectively

p_0	p_1	ℓ_1	ℓ_2	n_1	n	r_1	r	$\Pr(H_0 \mid r_1, n_1)$	$\Pr(H_1 \mid r_1, n_1)$	$\Pr(H_0 \mid r, n)$	$\Pr(H_1 \mid r, n)$
Prior: $p \sim \text{Beta}(1,1)$											
0.05	0.25	0.25	0.2	10	58	1	19	0.1019	0.1971	0.0000	0.9203
		0.3	0.22	8	48	1	16	0.0712	0.3003	0.0000	0.9164
		0.35	0.25	7	37	1	13	0.0572	0.3671	0.0000	0.9290
0.1	0.3	0.25	0.2	13	61	2	23	0.1584	0.1608	0.0000	0.9107
		0.3	0.22	10	51	2	20	0.0896	0.3127	0.0000	0.9283
		0.35	0.25	7	40	1	16	0.1869	0.2553	0.0000	0.9211
0.2	0.4	0.25	0.2	19	65	5	32	0.1958	0.1256	0.0000	0.9363
		0.3	0.22	14	54	4	27	0.1642	0.2173	0.0000	0.9339
		0.35	0.25	10	41	3	21	0.1611	0.2963	0.0000	0.9294
0.3	0.5	0.25	0.2	23	63	9	37	0.1528	0.1537	0.0000	0.9157
		0.3	0.22	16	51	7	31	0.1046	0.3145	0.0000	0.9368
		0.35	0.25	11	39	5	24	0.1178	0.3872	0.0000	0.9231
Prior: $p \sim \text{Beta}(8,12)$											
0.2	0.4	0.25	0.2	15	47	2	25	0.1254	0.0732	0.0000	0.9363
		0.3	0.22	15	36	2	20	0.1254	0.0732	0.0000	0.9339
		0.35	0.25	15	23	2	14	0.1254	0.0732	0.0000	0.9294
0.25	0.45	0.25	0.2	15	47	4	28	0.1193	0.0940	0.0000	0.9240
		0.3	0.22	15	35	4	22	0.1193	0.0940	0.0000	0.9222
		0.35	0.25	15	22	4	16	0.1193	0.0940	0.0000	0.9432
0.3	0.5	0.25	0.2	15	45	5	30	0.1929	0.0607	0.0000	0.9157
		0.3	0.22	15	33	5	24	0.1929	0.0607	0.0000	0.9368
		0.35	0.25	15	21	5	17	0.1929	0.0607	0.0000	0.9231

the target p_1, rather than focusing on the uninteresting null value p_0.

For the lower panel of Figure 1, we specify $\pi_2 = 0.9$ and $\ell_2 = p_1 - p_0$, and the plots represent the posterior distributions under the BET design with different specifications of (p_0, p_1) as the indifference region shrinks gradually. Clearly, the proposed BET design ensures that the majority of the posterior distribution lies beyond p_1, whereas under Simon's optimal design the mode of the posterior distribution always lies halfway between p_0 and p_1.

Moreover, we consider the use of a relatively informative prior for p and examine its impact on the design's operating characteristics. Based on the estimated success rate of around 40% in phase III trials (Gan et al., 2012), if we assume some exchangeability between phase II and phase III trial results, we may take the prior distribution to be Beta(8,12), which is equivalent to the information of 8 responses among 20 subjects (with a sample proportion of 40%). For such an informative prior distribution, the resultant first-stage sample size might be too small, and thus we set the minimum sample size in the first stage to be $n_{\min} = 15$ to allow more data to be accumulated before making a continuation decision. To be conservative, we only consider using such a prior distribution on hypotheses tests with $p_1 \geq 0.4$. The lower part of Table 2 shows the performances of the BET design under such an informative prior, which are more pessimistic and stringent than the case with a noninformative prior. Correspondingly, Figure 2 shows the plots of the prior density, likelihood and the posterior density under such an informative prior. As the prior mean is less than p_1, it is evident that under the influence of such a pessimistic prior distribution, the posterior distribution is shifted to the left of the likelihood function.

Further, we compare the operating characteristics of the BET design with those of two Bayesian phase II single-arm designs by Tan and Machin (2002), namely, the single threshold design (STD) and the dual threshold design (DTD). Both designs consist of two stages, and the decision boundaries at the end of each stage are based on the posterior probabilities. More specifically, let p_L and p_U denote two threshold response rates. At the end of the first stage, if $\Pr(p > p_U | y_1, n_1) > \gamma_1$, the trial would proceed into the second stage, otherwise it would stop for futility. At the end of the second stage, if $\Pr(p > p_U | y, n) > \gamma_2$, the drug would be declared as promising. The sample size n_1 is chosen as the smallest integer such that $\Pr(p > p_U | y_1 = (p_U + 0.05)n_1, n_1) > \gamma_1$, and n is chosen as the smallest integer such that $\Pr(p > p_U | y = (p_U + 0.05)n, n) > \gamma_2$. The DTD design has the same decision boundary and sample size as those of the STD design for the second stage, but its design parameters for the first

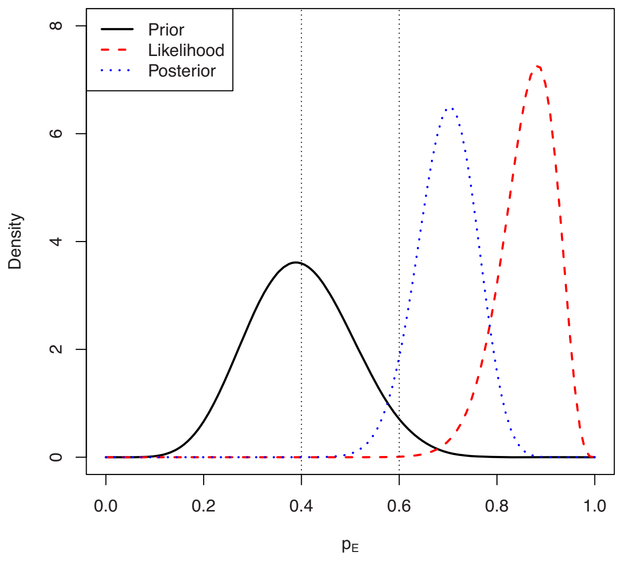

Figure 2. Comparison of the prior density, likelihood, and the posterior density at the end of the second stage when the response number reaches the minimum required levels under the Bayesian enhancement two-stage design for binary endpoint with $(p_0, p_1) = (0.4, 0.6)$, $(\ell_1, \ell_2) = (0.4, 0.2)$, and a prior distribution of Beta(8,12).

stage are different. The DTD design would stop for futility at the end of the first stage if $\Pr(p < p_L|y_1, n_1) > \gamma_1$, and its sample size for the first stage is calculated as the smallest integer satisfying $\Pr(p < p_L|y_1 = (p_L - 0.05)n_1, n_1) > \gamma_1$. In the original article, Tan and Machin (2002) recommended the settings of $(\gamma_1, \gamma_2) = (0.6, 0.7)$, $(0.6, 0.8)$, and (0.7,0.8). For a head-to-head comparison, we set $p_L = p_0$ and $p_U = p_1$, and compare the BET design with $(\pi_1, \pi_2) = (0.7, 0.8)$ against the STD and DTD designs with thresholds $(\gamma_1, \gamma_2) = (0.7, 0.8)$. Moreover, Tan and Machin (2002) recommended a prior distribution of Beta$(\eta + 1, 2 - \eta)$ for the response rate p where η is the prior mode; we adopt the same prior distribution by choosing $\eta = 0.1$.

Table 3 compares the operating characteristics of the BET, STD, and DTD designs, as well as Simon's two-stage design for various combinations of (p_0, p_1). We set $(\ell_1, \ell_2) = (0.15, 0.2)$ for the BET design and set $(\alpha, \beta) = (0.05, 0.05)$ for Simon's design, such that their maximum sample sizes are roughly similar to those of the STD and DTD designs. In terms of the posterior probabilities of H_0 and H_1 at the decision boundaries, it is evident that $\Pr(H_1|r, n)$, the posterior probabilities of H_1 at the second-stage decision boundaries, all exceed 0.8 for the three Bayesian designs, but are quite low for Simon's design. On the other hand, $\Pr(H_0|r_1, n_1)$, the posterior probabilities of H_0 at the first-stage decision boundaries are quite different across the three Bayesian designs: those under the STD design are the lowest and are consistently close to zero, those under the BET design remain below $1 - \pi_1 = 0.3$, and those under the DTD design exceed 0.5 and could be as high as 0.68. The reason for such high values of $\Pr(H_0|r_1, n_1)$ in the DTD design is that its first-stage decision boundary only requires the posterior probability $\Pr(p > p_0|r_1, n_1)$ to be greater than $1 - \gamma_1 = 0.3$, as the DTD design focuses on the stopping rule for futility regarding p_0, that is, the trial stops if $\Pr(p < p_0|r_1, n_1) > \gamma_1$. We compute the probabilities of early termination for futility at the end of the first stage (denoted by PET_0 and PET_1), and the expected sample sizes (denoted by ESS_0 and ESS_1) when the response rate equals p_0 and p_1, respectively. It appears that the STD design has much higher probabilities of early termination than the other designs and, as a result, its expected sample sizes under H_1 are quite low, indicating that the design might be overly stringent. For the two-stage designs, it is worth emphasizing that when the number of responders at the end of the first stage y_1 reaches the continuation criterion r_1, but is too small to result in any possible trial success in the second stage, that is, $y_1 + n - n_1 < r$, the trial should be terminated early despite the continuation criterion being met (Tan and Machin, 2006). None of the cases in Table 3 would have such a problem as they all satisfy $r_1 + n - n_1 > r$.

For the survival endpoint, we study the design characteristics under various specifications of the median survival times $(\theta_0, \theta_1) = (1, 2)$ and $(0.8, 1.8)$, and the lengths of HPD intervals $(\ell_1, \ell_2) = (0.4, 0.7)$ and $(0.5, 0.8)$. The cutoff values for the posterior probabilities are $(\pi_1, \pi_2) = (0.7, 0.8)$, and k_{HPD} is specified to be 1. We simulate failure times from Weibull distributions with shape parameter $k = 1$, 1.5, and 2. The median of the Weibull distribution is set to be $\theta_1 + 2$. We examine different censoring rates $c = 0.1$ and 0.2, and simulate the censoring times from a uniform distribution Unif$(0, L)$, where L is solved numerically such that $\Pr(T < C) = 1 - c$. We assume that the rate of accrual is 10 patients per time unit, and the interim monitoring starts after observing 10 failure events. Based on 1000 data replications, we compute the average sample sizes for the first stage and the entire trial, as well as the average posterior probabilities of H_0 and H_1 when a trial satisfies the continuation criterion in the first stage, denoted as $\Pr(H_0|\text{Stage 1})$ and $\Pr(H_1|\text{Stage 1})$, respectively, and when the drug is declared as promising in the second stage, denoted as $\Pr(H_0|\text{Stage 2})$ and $\Pr(H_1|\text{Stage 2})$, respectively.

Table 4 presents the design characteristics under various values of k, censoring rates c, and specifications of (θ_0, θ_1) and (ℓ_1, ℓ_2). The design is able to flexibly expand the sample size according to the censoring rate, as the average sample sizes are larger for higher values of c. A larger sample size is also associated with a shorter HPD interval. Moreover, when the HPD interval lengths (ℓ_1, ℓ_2) are fixed, the larger the values of (θ_0, θ_1), the larger the average sample size, because in relative terms the required HPD interval length becomes smaller for larger values of θ_0 and θ_1. The sample size appears to be insensitive to the value of k. As k is an unknown parameter, such a property makes the trial design more manageable, as it enables the investigators to gauge and adjust the average sample size at the design stage. The average posterior probability of H_0 is less than $1 - \pi_1$ when continuation is warranted in the first stage, and the average posterior probability of H_1 is more than π_2 when efficacy is declared in the second stage. This indicates that the design guarantees a high posterior probability of the alternative hypothesis being true when the drug is declared as promising at the end of the trial.

Table 3

Comparison of Bayesian enhancement two-stage (BET) design, single threshold design (STD), dual threshold design (DTD) and Simon's two-stage design, in terms of posterior probabilities of H_0 and H_1 at the decision boundaries, probabilities of early termination, and expected sample sizes

Design	p_0	p_1	n_1	n	r_1	r	$\Pr(H_0 \mid r_1, n_1)$	$\Pr(H_1 \mid r_1, n_1)$	$\Pr(H_0 \mid r, n)$	$\Pr(H_1 \mid r, n)$	PET_0	PET_1	ESS_0	ESS_1
BET	0.1	0.3	13	63	2	22	0.165	0.140	0.000	0.801	0.621	0.064	31.9	59.8
STD			24	61	9	22	0.000	0.783	0.000	0.845	1.000	0.725	24.0	34.2
DTD			27	61	2	22	0.540	0.003	0.000	0.845	0.233	0.001	53.1	61.0
Simon			20	49	3	9	0.157	0.075	0.025	0.036	0.677	0.035	29.4	48.0
BET	0.15	0.35	15	67	3	27	0.226	0.113	0.000	0.809	0.604	0.062	35.6	63.8
STD			30	70	12	28	0.000	0.702	0.000	0.801	0.999	0.655	30.0	43.8
DTD			29	70	3	28	0.685	0.002	0.000	0.801	0.168	0.001	63.1	70.0
Simon			28	60	6	14	0.142	0.064	0.036	0.026	0.765	0.039	35.5	58.7
BET	0.2	0.4	19	70	5	32	0.216	0.104	0.000	0.820	0.673	0.070	35.7	66.4
STD			35	78	16	36	0.000	0.729	0.000	0.853	0.999	0.700	35.0	47.9
DTD			29	78	5	36	0.589	0.005	0.000	0.853	0.284	0.002	64.1	77.9
Simon			28	62	7	18	0.226	0.047	0.040	0.035	0.678	0.031	38.9	60.9
BET	0.25	0.45	20	70	6	36	0.284	0.077	0.000	0.840	0.617	0.055	39.1	67.2
STD			40	84	20	42	0.000	0.702	0.000	0.801	0.999	0.684	40.0	53.9
DTD			29	84	6	42	0.673	0.003	0.000	0.801	0.232	0.002	71.3	83.9
Simon			30	70	9	24	0.251	0.042	0.040	0.031	0.674	0.031	43.1	68.8
BET	0.3	0.5	22	69	8	39	0.261	0.082	0.000	0.835	0.671	0.067	37.5	65.9
STD			44	88	25	49	0.000	0.778	0.000	0.834	1.000	0.774	44.0	53.9
DTD			27	88	7	49	0.664	0.005	0.000	0.834	0.256	0.003	72.4	87.8
Simon			34	71	12	28	0.252	0.035	0.046	0.032	0.693	0.029	45.4	69.9

4. Trial Applications

4.1. *Advanced Hepatocellular Cancer Trial*

Back to the phase II study on erlotinib in patients with advanced hepatocellular cancer, the outcome of interest was the proportion of progression-free patients at 24 weeks (Philip et al., 2005). Simon's optimal two-stage design was used to calibrate the design parameter of this phase II study, with $(p_0, p_1) = (0.05, 0.2)$ under the type I and type II error rate constraints $(\alpha, \beta) = (0.09, 0.08)$. The optimal design parameters under such specifications are $(n_1, n, r_1, r) = (15, 35, 1, 4)$. The trial enrolled three additional patients beyond the originally planned 35 subjects, and ended with trial success (12 responses among 38 subjects) and continuation into a phase III study of 720 patients, which however concluded with failure eventually (Zhu et al., 2015).

As an illustration, we apply the BET design to such a phase II trial. During the design stage, we specify the minimum required posterior probability to be $\pi_1 = \pi_2 = 1 - \alpha$, and we may set $\ell_2 = 0.25$ and $\ell_1 = 0.4$, which would result in the design parameters of $(n_1, n, r_1, r) = (8, 35, 1, 10)$. The design would require eight patients in the first stage and 27 patients in the second stage, and a total of 10 responses are necessary for declaring the drug promising. The total sample size of the proposed design would be same as that of Simon's optimal two-stage design. The trial enrolled three additional patients beyond the originally planned sample size in Simon's design. To be conservative, we assume that these patients are responders and remove them from the total number of responses. Had the trial been conducted under the BET design, the drug would not have been declared as promising and would not have proceeded into the phase III study.

4.2. *Soft Tissue Sarcoma Trial*

Patel et al. (1997) reported a phase II study on the effects of paclitaxel (taxol) in the treatment of soft tissue sarcoma. The study was conducted based on Simon's optimal two-stage design with $(p_0, p_1) = (0.05, 0.2)$ under the type I and type II error rate constraints $(\alpha, \beta) = (0.1, 0.1)$. The optimal design parameters under such specifications are $(n_1, n, r_1, r) = (14, 37, 1, 3)$. The trial was terminated early for futility at the end of the first stage, as no response had been observed in the first 14 subjects. Had the trial been conducted under the BET design with the same first-stage sample size, it would also have been stopped early for futility as the threshold of the BET design is more stringent than that of Simon's design, and thus the same conclusion as the original design would have been drawn.

As another illustration, we apply the BET design to such a phase II trial. In the trial design, we specify the minimum required posterior probability to be $\pi_1 = \pi_2 = 0.9$, and require the desirable length of the HPD interval to be $\ell_2 = p_1 - p_0 = 0.15$ and $\ell_1 = 2\ell_2 = 0.3$. With such specifications, the parameters of our Bayesian two-stage design would be $(n_1, n, r_1, r) = (14, 90, 2, 23)$. The stage 1 sample size $n_1 = 14$ is the same for Simon's and our designs, while the stage 1 cutoff values r_1 are different: $r_1 = 2$ in our design but $r_1 = 1$ in Simon's design. Our total sample size is much larger than Simon's, while it can be reduced if we take a large value of ℓ_2. At the boundary parameters, $\hat{p} = r/n = 3/37 \approx 0.08$

Table 4

Bayesian enhancement two-stage designs for survival endpoint with $(\pi_1, \pi_2) = (0.7, 0.8)$ *under various values of the parameter* k *in the Weibull distribution and censoring rates* c*, specifications of* (ℓ_1, ℓ_2) *and* (θ_0, θ_1)*, the average sample sizes for the first stage and the whole trial (denoted as* $\bar{n}_1$ *and* $\bar{n}$*), and the average posterior probabilities of* H_0 *and* H_1 *when a trial achieves success in the first stage, and when the drug is declared as promising in the second stage, respectively*

k	c	θ_0	θ_1	ℓ_1	ℓ_2	$\bar{n}_1$	$\bar{n}$	$\Pr(H_0 \mid \text{Stage 1})$	$\Pr(H_1 \mid \text{Stage 1})$	$\Pr(H_0 \mid \text{Stage 2})$	$\Pr(H_1 \mid \text{Stage 2})$
1	0.1	1	2	0.5	0.8	23.2	57.7	0.0660	0.1437	0.0000	0.9127
				0.4	0.7	35.4	73.1	0.0255	0.2572	0.0000	0.9346
		0.8	1.8	0.5	0.8	15.3	47.5	0.0719	0.1063	0.0000	0.9148
				0.4	0.7	23.2	60.9	0.0353	0.2135	0.0000	0.9331
	0.2	1	2	0.5	0.8	25.6	63.9	0.0503	0.1932	0.0000	0.9275
				0.4	0.7	39.1	81.1	0.0131	0.3689	0.0000	0.9531
		0.8	1.8	0.5	0.8	16.9	52.8	0.0571	0.1844	0.0000	0.9348
				0.4	0.7	25.4	67.4	0.0220	0.2774	0.0000	0.9505
1.5	0.1	1	2	0.5	0.8	24.0	58.8	0.0024	0.6191	0.0000	0.9893
				0.4	0.7	36.4	74.2	0.0000	0.8669	0.0000	0.9979
		0.8	1.8	0.5	0.8	16.1	48.5	0.0052	0.5409	0.0000	0.9871
				0.4	0.7	24.0	62.1	0.0006	0.7303	0.0000	0.9967
	0.2	1	2	0.5	0.8	27.7	66.6	0.0013	0.7447	0.0000	0.9958
				0.4	0.7	41.3	84.0	0.0000	0.9106	0.0000	0.9994
		0.8	1.8	0.5	0.8	18.6	55.0	0.0028	0.6009	0.0000	0.9938
				0.4	0.7	27.3	69.6	0.0002	0.7913	0.0000	0.9988
2	0.1	1	2	0.5	0.8	24.6	59.1	0.0000	0.9381	0.0000	0.9999
				0.4	0.7	36.8	74.6	0.0000	0.9895	0.0000	0.9999
		0.8	1.8	0.5	0.8	16.7	49.1	0.0002	0.8594	0.0000	0.9998
				0.4	0.7	24.6	62.4	0.0000	0.9676	0.0000	0.9999
	0.2	1	2	0.5	0.8	28.5	67.2	0.0000	0.9686	0.0000	0.9999
				0.4	0.7	42.2	84.6	0.0000	0.9963	0.0000	0.9999
		0.8	1.8	0.5	0.8	19.6	56.0	0.0001	0.8981	0.0000	0.9999
				0.4	0.7	28.5	71.2	0.0000	0.9823	0.0000	0.9999

under Simon's design, but $\hat{p} = r/n = 23/90 \approx 0.26$ using our design. Considering $p_1 = 0.2$ as the minimum efficacy level for the experimental drug, it is obvious that Simon's estimated response rate does not achieve this minimum level.

5. Discussion

Based on posterior probabilities, the BET design is more natural and ensures that the drug's response rate reaches the desirable target level, while using the clinically uninteresting level as the continuation criterion. We consider it necessary to choose a target level that exceeds the uninteresting rate by a certain margin, rather than targeting the uninteresting rate as Simon's design does. In that sense, the proposed design is more stringent than Simon's design, that is, if our design claims the experimental drug promising, it is more likely to be true and thus would reduce the failure possibility of the subsequent phase III trial. The BET design, which utilizes posterior probabilities and HPD interval lengths for decision making, is flexible and can be easily adapted to accommodate different types of endpoints and prior distributions.

Joseph et al. (1995) and M'Lan et al. (2008) proposed several Bayesian sample size criteria for binomial proportions. Based on the prior predictive distribution of the number of responses, which follows a beta–binomial distribution, we may obtain a series of HPD interval lengths corresponding to different numbers of responses. The sample size criteria can be constructed based on the average, median, or the maximum these HPD interval lengths. Their approaches take into account the randomness of the data, while our posterior probabilities are calculated under the assumption that the smallest number of responses necessary for continuation or for declaring the trial success is observed without considering the random nature of the data.

6. Supplementary Materials

The software for implementing the BET designs for binary and survival endpoints is available with this article at the *Biometrics* website on Wiley Online Library.

Acknowledgement

We thank the associate editor, the referees, and the editor for their many constructive and insightful comments that have led to significant improvements in the article. The research was supported by a grant (17326316) from the Research Grants Council of Hong Kong.

References

Baey, C. and Le Deley, M. C. (2011). Effect of a misspecification of response rates on type I and type II errors, in a phase II Simon design. *European Journal of Cancer* **47**, 1647–1652.

Chen, K. and Shan, M. (2008). Optimal and minimax three-stage designs for phase II oncology clinical trials. *Contemporary Clinical Trials* **29**, 32–41.

Ensign, L. G., Gehan, E. A., Kamen, D. S., and Thall, P. F. (1994). An optimal three-stage design for phase II clinical trials. *Statistics in Medicine* **13**, 1727–1736.

Gan, H. K., You, B., Pond, G. R., and Chen, E. X. (2012). Assumptions of expected benefits in randomized phase III trials evaluating systemic treatments for cancer. *Journal of National Cancer Institute* **104**, 590–598.

Green, S. J. and Dahlberg, S. (1992). Planned versus attained design in phase II clinical trials. *Statistics in Medicine* **11**, 853–862.

Joseph, L., Wolfson, D. B., and Berger, R. (1995). Sample size calculations for binomial proportions via highest posterior density intervals. *The Statistician* **44**, 143–154.

Koyama, T. and Chen, H. (2008). Proper inference from Simon's two-stage designs. *Statistics in Medicine* **27**, 3145–3154.

Kola, I. and Landis, J. (2004). Can the pharmaceutical industry reduce attrition rates? *Nature Reviews Drug Discovery* **3**, 711–716.

Lee, J. J. and Feng, L. (2005). Randomised phase II designs in cancer clinical trials: Current status and future directions. *Journal of Clinical Oncology* **23**, 4450–4457.

Liu, J. F., Lin, Y., and Shih. W. J. (2010). On Simon's two-stage design for single-arm phase IIA cancer clinical trials under beta–binomial distribution. *Statistics in Medicine* **29**, 1084–1095.

Lee, J. J. and Liu, D. D. (2008). A predictive probability design for phase II cancer clinical trials. *Clinical Trials* **5**, 93–106.

M'Lan, C.E., Joseph, L., and Wolfson, D.B. (2008). Bayesian sample size determination for binomial proportions. *Bayesian Analysis* **3**, 269–296.

Mander, A. P. and Thompson, S. G. (2010). Two-stage designs optimal under the alternative hypothesis for phase II cancer clinical trials. *Contemporary Clinical Trials* **31**, 572–578.

Patel, S., Linke, K., Burgess, A., Papadopoulos, N., Plager, C., Jenkins, J., and Benjamin, R. (1997). Phase II study of paclitaxel in patients with soft tissue sarcomas. *Sarcoma* **1**, 95–97.

Philip, P. A. and Mahoney, M. R., Allmer, C., Thomas, J., Pitot, H. C., Kim, G., Donehower, R. C., Fitch, T., Picus, J., and Erlichman, C. (2005). Phase II study of Erlotinib (OSI-774) in patients with advanced hepatocellular cancer. *Journal of Clinical Oncology* **23**, 6657–6663.

Retzios, A. D. (2009). *Why Do So Many Phase 3 Clinical Trials Fail?* San Ramon, CA: Bay Clinical R&D Services.

Sambucini, V. (2008). A Bayesian predictive two-stage design for phase II clinical trials. *Statistics in Medicine* **27**, 1199–1224.

Sambucini, V. (2010). A Bayesian predictive strategy for an adaptive two-stage design in phase II clinical trials. *Statistics in Medicine* **29**, 1430–1442.

Shan, G., Wilding, G. E., Hutson, A. D., and Gerstenberger, S. (2016). Optimal adaptive two-stage designs for early phase II clinical trials. *Statistics in Medicine* **35**, 1257–1266.

Shuster, J. (2002). Optimal two-stage designs for single arm phase II cancer trials. *Journal of Biopharmaceutical Statistics* **12**, 39–51.

Simon, R. (1989). Optimal two-stage designs for phase II clinical trials. *Controlled Clinical Trials* **10**, 1–10.

Thall, P. F. and Simon, R. (1994). Practical Bayesian guidelines for phase IIb clinical trials. *Biometrics* **50**, 337–349.

Tan, S. B. and Machin, D. (2002). Bayesian two-stage designs for phase II clinical trials. *Statistics in Medicine* **21**, 1991–2012.

Tan, S. B. and Machin, D. (2006). Letter to the editor: Bayesian two-stage designs for phase II clinical trials. *Statistics in Medicine* **25**, 3407–3408.

Yin, G. (2012). *Clinical Trial Design: Bayesian and Frequentist Adaptive Methods*, Hoboken, NJ: John Wiley & Sons.

Zhu, A. X., Rosmorduc, O., Evans, T. R. J., Ross, P. J., Santoro, A., Carrilho, F. J, et al. (2015). SEARCH: A phase III, randomized, double-blind, placebo-controlled trial of Sorafenib plus Erlotinib in patients with advanced hepatocellular carcinoma. *Journal of Clinical Oncology* **33**, 559–566.

Received June 2017. *Revised January* 2018.
Accepted January 2018.